RNA-Seq on non-model organisms

Jorge Espinosa[2], Jordi Durban[1], Juan J. Calvete[1]
1. Laboratori Venòmica Evolutiva i Traslacional, Institut Biomedicina València (CSIC)
2. Departament de Genètica, Universitat de València
email: jdurban@ibv.csic.es

Background

Venoms research has been continuously enhanced by advances in technology. In particular, the emergence of 'omic' technologies at the turn of the twenty-first century has revolutionized biological research. The breakthrough experienced by venom research in the last decade is due to the development and application of omics technology to the qualitative and quantitative profiling of the venom gland mRNA expression (venom gland transcriptomics) and the precise identification of the components expressed in the venom (venomics). Both approaches have been fueled by advances on sequencing technologies and improvements on **RNASeq**. However, in the absence of a genomic reference, several approaches have been developed in order to analyze mRNA expression data from Next Generation Sequencing technologies which are not as straightforward as those developed when a genome reference sequence is available.

de novo RNA-Seq assembly

Quality check

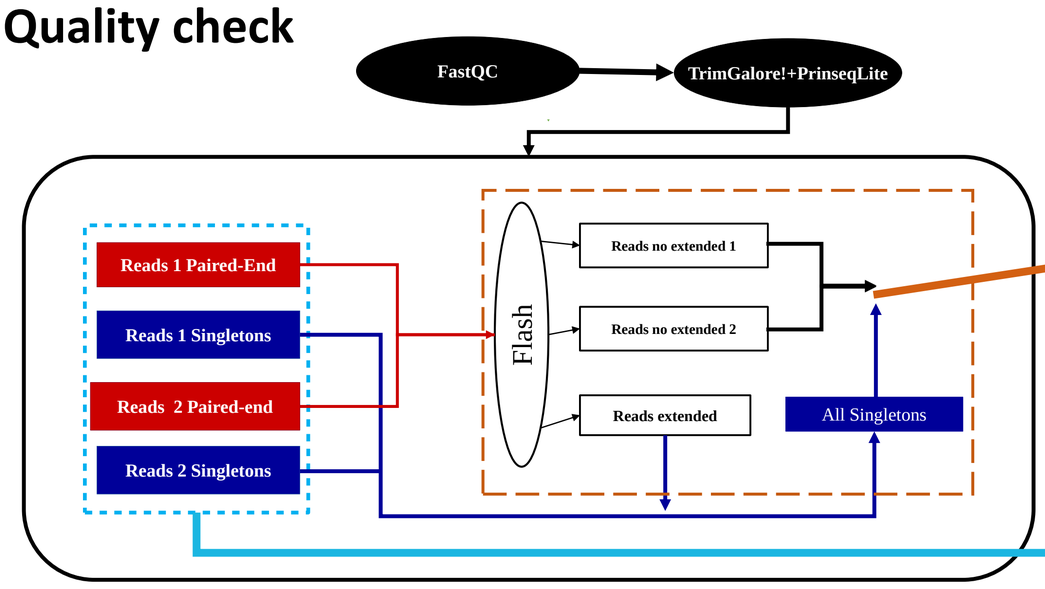
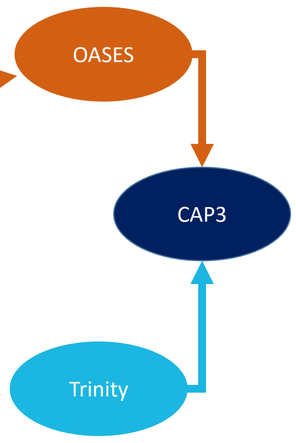


Figure1. Overview Illumina preprocessing procedure

Trimming and adapter removing
- Reads with a quality score below 30, adapters, and the first 13 bases were eliminated using FastQC, TrimGalore! and PRINSEQ.
- Overlapping short paired-end reads were extended with Fast Length Adjustment of SHort reads tool.

FASTQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trim Galore: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
PRINSEQ: http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi
FLASH: https://ccb.jhu.edu/software/FLASH/

Assembly



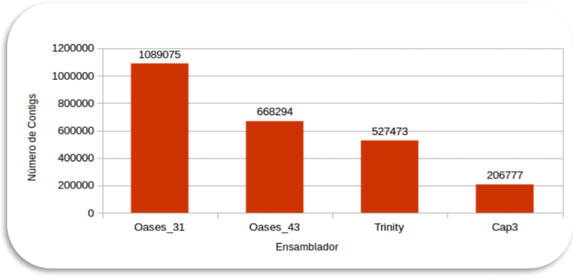


Figure 2. Number of contigs by assembler

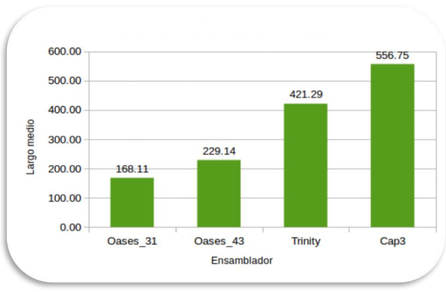


Figure 3. Average contig length by assembler (nt)

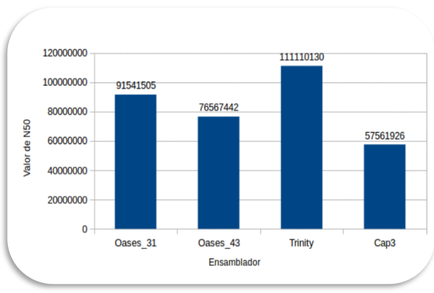


Figure 4. N50 by assembler (nt)

Multi – assembler approach
- Oases (Velvet) is an assembler that uses De Bruijn graph (DBG) to generate contigs.
- Trinity also uses DBG to generate contigs.
- Cap3 uses and Overlapping Layout Consensus (OLC) algorithm.

Oases: https://www.ebi.ac.uk/~zerbino/oases/
Trinity: https://github.com/trinityrnaseq/trinityrnaseq/wiki
CAP3: http://seq.cs.iastate.edu/cap3.html

DBG: Works better with a great number of small reads.
OLC: Works better with few cured long reads

Annotation

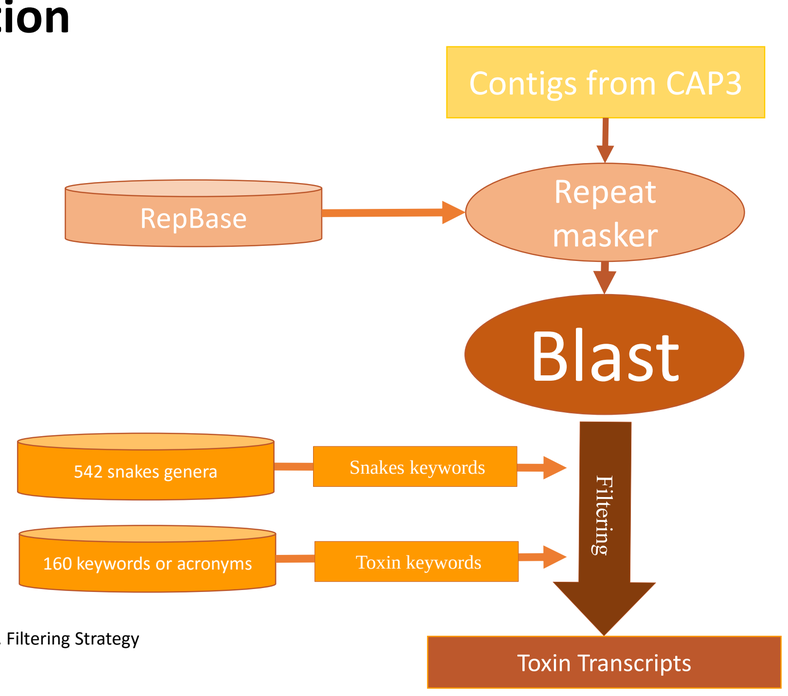
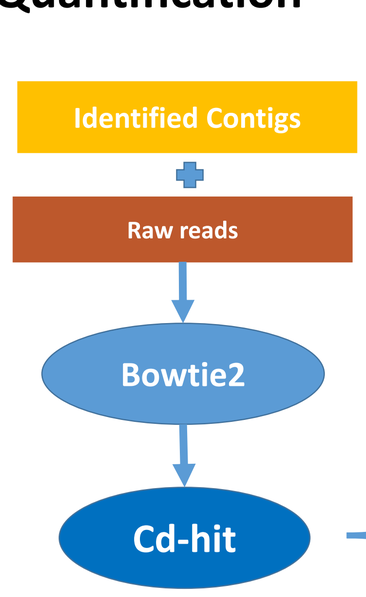


Figure . Filtering Strategy

Homology search annotation process
- Transcript obtained by the assemblers were filtered to eliminate repetitive elements.
- The we used Biopython to extract the information needed of each hit.
- Then each informative hit was filltered by the snakes and toxin keywords

Blast: https://blast.ncbi.nlm.nih.gov
Biopython: https://biopython.org/wiki/Download

Quantification



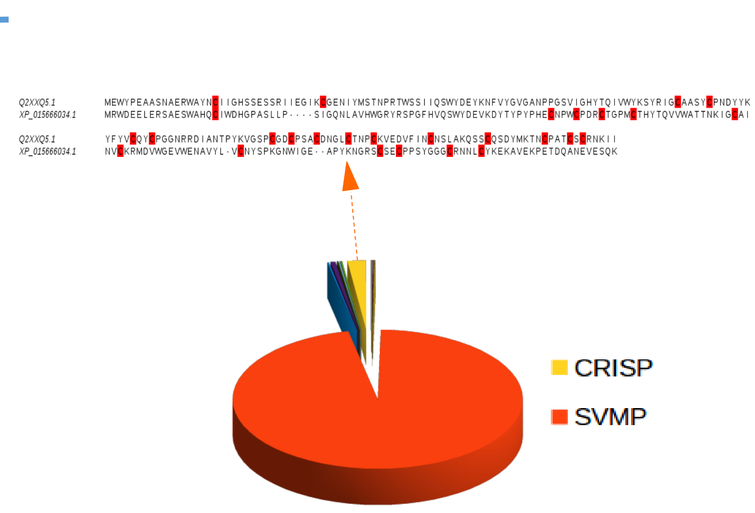


Figure . Pie Chart of the relative abundance if the major toxin families in the venom gland transcriptome.

Quantification of transcript abundance
- Transcript abundance was estimated by mapping paired and unpaired trimmed reads back to annotated encoding transcripts with Bowtie2 and expressed in RPKM.
- Molecular protein diversity was assessed by clustering translated sequences with cd-hit

Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Cd-hit: http://weizhongli-lab.org/cd-hit/

Concluding remarks

Transcriptomics creates many challenges for scientists who are unfamiliar with working with RNA, managing large NGS datasets and executing the required programs, specially if no reference sequence is available. This work provides the methodology used at the *Evolutionary and translational venomics* lab for the bioinformatic analysis of targeted sequencing of transcripts from snake venom glands. The species-specific obtained transcriptome greatly improves the venom proteomic profiling which takes advantage from a custom protein database.